

ウェブから収集した専門分野コーパスと要素合成法を用いた 専門用語訳語推定

外池 昌嗣[†]・宇津呂武仁^{††}・佐藤 理史^{†††}

本論文では、ウェブを利用した専門用語の訳語推定法について述べる。これまでに行われてきた訳語推定の方法の1つに、パラレルコーパス・コンパラブルコーパスを用いた訳語推定法があるが、既存のコーパスが利用できる分野は極めて限られている。そこで、本論文では、訳を知りたい用語を構成する単語・形態素の訳語を既存の対訳辞書から求め、これらを結合することにより訳語候補を生成し、単言語コーパスを用いて訳語候補を検証するという手法を採用する。しかしながら、単言語コーパスであっても、研究利用可能なコーパスが整備されている分野は限られている。このため、本論文では、ウェブをコーパスとして用いる。ウェブを訳語候補の検証に利用する場合、サーチエンジンを通してウェブ全体を利用する方法と、訳語推定の前にあらかじめ、ウェブから専門分野コーパスを収集しておく方法が考えられる。本論文では、評価実験を通して、この2つのアプローチを比較し、その得失を論じる。また、訳語候補のスコア関数として多様な関数を定式化し、訳語推定の性能との間の相関を評価する。実験の結果、ウェブから収集した専門分野コーパスを用いた場合、ウェブ全体を用いるよりカバーレージは低くなるが、その分野の文書のみを利用して訳語候補の検証を行うため、誤った訳語候補の生成を抑える効果が確認され、高い精度を達成できることがわかった。

キーワード：要素合成法、訳語推定、ウェブ、専門用語

Compositional Translation Estimation of Technical Terms Using a Domain/Topic-Specific Corpus Collected from the Web

MASATSUGU TONOIKE[†], TAKEHITO UTSURO^{††} and SATOSHI SATO^{†††}

This paper studies how to compile a bilingual lexicon for technical terms using the Web. In the task of estimating bilingual term correspondences of technical terms, it is usually rather difficult to find an existing corpus for the domain of such technical terms. In this paper, we adopt an approach of collecting a corpus for the domain of such technical terms from the Web. As a method of translation estimation for technical terms, we employ a compositional translation estimation technique, where translation candidates of a term are compositionally generated by concatenating the

[†] 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University

^{††} 筑波大学大学院システム情報工学研究科, Graduate School of Systems and Information Engineering, University of Tsukuba

^{†††} 名古屋大学大学院工学研究科, Graduate School of Engineering, Nagoya University

translation of the constituents of the term. Then, the generated translation candidates are validated using the domain/topic-specific corpus collected from the Web. This paper further quantitatively compares the proposed approach with another approach of validating translation candidates directly through a search engine. We show that the domain/topic-specific corpus collected from the Web contributes to achieving higher precision in translation candidate validation.

Key Words: *Compositional translation estimation, Web, Echnical term*

1 はじめに

本論文では、ウェブを利用した専門用語の訳語推定法について述べる。専門用語の訳語情報は、技術翻訳や同時通訳、機械翻訳の辞書の強化などの場面において、実に様々な分野で求められている。しかしながら、汎用の対訳辞書には専門用語がカバーされていないことが多い、対訳集などの専門用語の訳語情報が整備されている分野も限られている。その上、専門用語の訳語情報が整備されていたとしても、最新の用語を追加していく作業が必要になる。このため、あらゆる分野で、専門用語の訳語情報を人手で整備しようとすると、大変なコストとなる。そこで、本論文では、対象言語を英語、日本語双方向とし、自動的に専門用語の訳語推定を行う方法を提案する。

これまでに行われてきた訳語推定の方法の1つに、パラレルコーパスを用いた訳語推定法がある (Matsumoto and Utsuro 2000)。しかしながら、パラレルコーパスが利用できる分野は極めて限られている。これに対して、対訳関係のない同一分野の2つの言語の文書を組にしたコンパラブルコーパスを利用する方法 (Fung and Yee 1998; Rapp 1999) が研究されている。これらの手法では、コーパスにそれぞれ存在する2言語の用語の組に対して、各用語の周囲の文脈の類似性を言語を横断して測定することにより、訳語対応の推定が行われる。パラレルコーパスに比べればコンパラブルコーパスは収集が容易であるが、訳語候補が膨大となるため、精度の面で問題がある。また、この方法では、訳語推定対象の用語を構成する単語・形態素の情報を利用していない。これに対して、(藤井, 石川 2000; Baldwin and Tanaka 2004) では、訳を知りたい用語を構成する単語・形態素の訳語を既存の対訳辞書から求め、これらを結合することにより訳語候補を生成し、単言語コーパスを用いて訳語候補を検証するという手法を提案している。(以下、本論文では、用語の構成要素の訳語を既存の対訳辞書から求め、これらを結合することにより訳語候補を生成する方法を「要素合成法」と呼ぶ。)

要素合成法による訳語推定法の有効性を調査するために、既存の専門用語対訳辞書の10分野から、日本語と英語の専門用語で構成される訳語対を617個抽出した¹。そして、それぞれの訳

¹ 4.1節で述べる未知訳語対集合 Y_{ST} に対応する。

語対の日本側の用語と英語側の用語の構成要素が対応しているかを調べたところ、88.5%の訳語対で日英の構成要素が対応しているという結果が得られた。このことから、専門用語に対して要素合成法による訳語推定法を適用することは有効である可能性が高いことがわかった。（以下、本論文では、訳語対において各言語の用語の構成要素が対応していることを「構成的」と呼ぶものとする。）

しかしながら、単言語コーパスであっても、研究利用可能なコーパスが整備されている分野は限られている。このため、本論文では、大規模かつあらゆる分野の文書を含むウェブをコーパスとして用いるものとする。ウェブを訳語候補の検証に利用する場合、(Cao and Li 2002) の様に、サーチエンジンを通してウェブ全体を利用して訳語候補の検証を行うという方法がまず考えられる。その対極にある方法として、訳語推定の前にあらかじめ、ウェブから専門分野コーパスを収集しておくことも考えられる。サーチエンジンを通してウェブ全体を利用するアプローチは、カバレージに優れるが、様々な分野の文書が含まれるため誤った訳語候補を生成してしまう恐れもある。また、それぞれの訳語候補に対してサーチエンジンで検索を行わなければいけないため、サーチエンジン検索の待ち時間が無視できない。これに対して、ウェブから専門分野コーパスを収集するアプローチは、ウェブ全体を用いるよりカバレージは低くなるが、その分野の文書のみを利用して訳語候補の検証を行うため、誤った訳語候補を削除する効果が期待できる。また、ひとたび専門分野コーパスを収集すれば、訳語推定対象の用語が大量にある場合でも、サーチエンジンを介してウェブにアクセスすることなく訳語推定を行うことができる。しかしながら、これまで、この2つのアプローチの比較は行われてこなかったため、本論文では、評価実験を通して、この2つのアプローチを比較し、その得失を論じる。

さらに、上記の2つのアプローチの比較も含めて、本論文では、訳語候補のスコア関数として、多様な関数を以下のように定式化する。要素合成法では、構成要素に対して、対訳辞書中の訳語を結合することにより訳語候補が生成されるので、構成要素の訳語にもとづいて訳語候補の適切さを評価する。これを対訳辞書スコアと呼ぶ。また、それとは別に、生成された訳語候補がコーパスに生起する頻度に基づいて、訳語候補の適切さを評価する。これをコーパススコアと呼ぶ。本論文では、この2つスコアの積で訳語候補のスコアを定義する。本論文では、対訳辞書スコアに頻度と構成要素長を考慮したスコアを用い、また、コーパススコアには頻度に基づくスコアを用いたスコア関数を提案し、確率に基づくスコア関数（藤井、石川 2000）と比較する。さらに、対訳辞書スコア、コーパススコアとしてどのような尺度を用いるか、に加え、訳語候補の枝刈りにスコアを使うかどうか、コーパスとしてウェブ全体を用いるか専門分野コーパスを用いるか、といったスコア関数の設定を変化させて合計12種類のスコア関数を定義し、訳語推定の性能との間の相関を評価する。

実験の結果、コーパスとしてウェブ全体を用いた場合、ウェブには様々な分野の文書が含まれるため誤った訳語候補を生成してしまうことが多い反面、カバレージに優れることがわかつ

た。逆に、ウェブから収集した専門分野コーパスを用いた場合、ウェブ全体を用いるよりカバレージは低くなるが、その分野の文書のみを利用して訳語候補の検証を行うため、誤った訳語候補の生成を抑える効果が確認された。また、ウェブから収集した専門分野コーパスを用いる方法の性能向上させるためには、専門分野コーパスに含まれる正解訳語の割合を改善することが課題であることがわかった。

以下、本論文では、第2章でウェブを用いた専門用語訳語推定の枠組みを導入し、専門分野コーパスの収集方法について述べる。第3章では要素合成法による訳語推定の定式化を行い、訳語候補のスコア関数を導入する。第4章では実験と評価について述べる。第5章では関連研究について述べ、本論文との相違点を論じる。

2 ウェブを用いた専門用語訳語推定

2.1 概要

ウェブを用いた専門用語の訳語推定の全体像を図1に示す。本論文では、言語 S の専門用語が複数個が与えられたとき、それらの用語に対して、言語 T における訳語を推定するという問題を考える。このような状況としては、例えば、ある専門分野において、まとまった数の専門文書が与えられ、それらの文書から用語を抽出し、専門用語の対訳辞書を作成する場合を考えられる。あるいは、ある専門分野の文書と既存の汎用対訳辞書があり、この文書を翻訳家が翻訳したい場合などが考えられる。ここで、一般に、与えられた複数の専門用語は、既存の汎用対訳辞書に含まれる訳語の個数にしたがって、訳語が1個である用語の集合 X_S^U 、訳語が2個

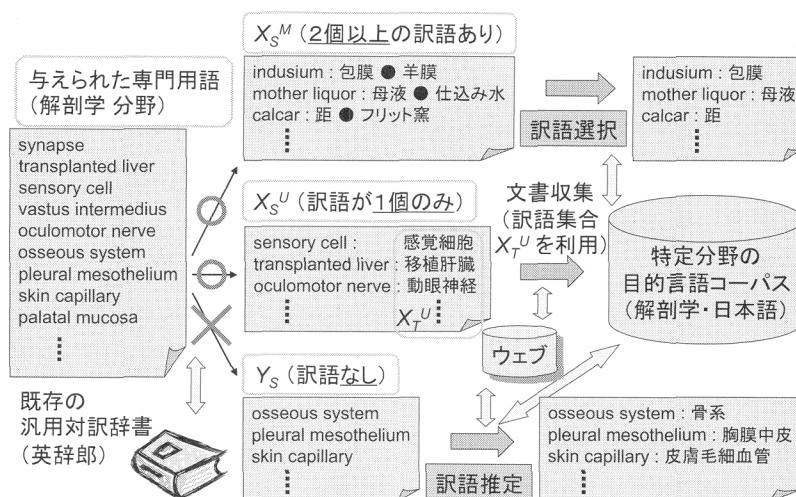


図1 ウェブを用いた専門用語訳語推定

以上である用語の集合 X_S^M , そして, 訳語が得られない用語の集合 Y_S という 3 つの部分集合に分けられる. 本論文では, 既存の辞書に訳語が 1 個だけ含まれる用語の集合 X_S^U の訳語は正しいと仮定し, 集合 X_S^U の用語の訳語の集合 X_T^U を用いてウェブから専門分野コーパスを収集し, 訳語推定に利用するものとする.

本論文では, 既存の対訳辞書で訳語が得られない用語の集合 Y_S を訳語推定の対象とする. 一方, 集合 X_S^M の用語に対しては, 既存の対訳辞書にある訳語の中から最も適切なものを選択する必要がある. 例えば, 論理回路分野に属する日本語の専門用語「レジスタ」の訳語としては, サッカー用語の “regista” ではなく, 用語 “register” が選択されなければならない. この訳語選択の課題については, (Tonoike, Kida, Takagi, Sasaki, Utsuro, and Sato 2005)において, すでに一定の成果が得られており, ウェブから収集した専門分野コーパスに生起する頻度の最も大きい訳語を選択することにより, 英日方向で 69%, 日英方向で 75% の正解率が得られたと報告されている. そこで, 本論文では, 集合 X_S^M の用語の訳語選択の課題は取り扱わない.

2.2 専門分野コーパスの収集

本論文では, 言語 T の専門分野コーパスをウェブから収集して訳語推定に利用する. この専門分野コーパスを集める際には, 既存対訳辞書に訳語が 1 つだけ存在する専門用語の訳語の集合 X_T^U を利用する. 具体的には, 集合 X_T^U に含まれる用語 x_T^U を含むサーチエンジンのクエリーを用いてウェブから上位 100 ページを収集する². それらのページに, 用語 x_T^U がアンカーテキストとなっているアンカーが存在する場合は, そのアンカー先ページも入手する. これを, 集合 X_T^U に含まれる用語すべてに対して行い, 収集されたウェブページを集めて, 専門分野コーパスとする. 日本語のコーパスを収集する際に用いたクエリーは, “ x_T^U とは”, “ x_T^U という”, “ x_T^U は”, “ x_T^U の”, 及び, “ x_T^U ” である. 一方, 英語のコーパスを収集する際に用いたクエリは一, “ x_T^U AND what's”, “ x_T^U AND glossary”, 及び, “ x_T^U ” である. ここでは, 専門用語 x_T^U について記述されている文書, 例えば, オンライン用語集などを上位にランクするために, 経験的にこれらクエリーを用いている.

ここで, 計算コストや記憶容量の問題を考慮した上で, できるだけ訳語推定の性能を向上させるためには, 訳語推定対象の分野の用語を十分に含むできるだけ小さいコーパスを収集することが望ましい. これを実現する方法として, サーチエンジンのクエリーに複数の用語を含めたり, 取得すべきページ数を変更することなどが考えられる. しかしながら, 本論文では, ウェブから収集した専門分野コーパスを利用する方式と, サーチエンジンを通してウェブ全体を利用する方式の比較に焦点を当てるため, 訳語推定対象の分野の用語を十分に含むできるだけ小さいコーパスを収集する方式を確立することは, 論文の対象外とする. この問題に関連する知

² 本論文では, サーチエンジンを用いる場合, 日本語のクエリーの場合は goo (<http://www.goo.ne.jp>) を用い, 英語のクエリーの場合は Yahoo! (<http://www.yahoo.com/>) を用いる.

見としては、(高木, 木田, 外池, 佐々木, 日野, 宇津呂, 佐藤 2005) の研究がある。(高木他 2005) では、評価用の正解訳語の集合を設定し、上記の方法によってウェブから収集したコーパスに、評価用の正解訳語を含む割合の評価を行った。その結果、サーチエンジンのヒット数に上限を設けてコーパス収集に使用する用語の数を絞り込むことにより、コーパス収集に使用する用語の数を少なくしても、評価用の正解訳語を含む割合が下がらないことが報告されている。また、(高木他 2005) では、訳語推定対象の用語（および評価用の正解訳語）とは異なる分野のコーパスを利用する場合についても評価を行っている。これによると、評価用の正解訳語を含む割合は、訳語推定対象の用語と近い分野のコーパスを用いた場合は低下しないが、全く異なった分野のコーパスを用いた場合は低下することを実験的に確かめている。このことは、訳語推定対象の用語の分野となるべく近い分野のコーパスを用いて訳語推定をすべきであることを示している。

3 要素合成法による専門用語訳語推定の定式化

3.1 概要

要素合成法による訳語推定の例として、日本語の専門用語「応用行動分析」の訳語推定の様子を図 2 に示す。まず、既存の対訳辞書を参照し、日本語の見出しを検索することにより、日本

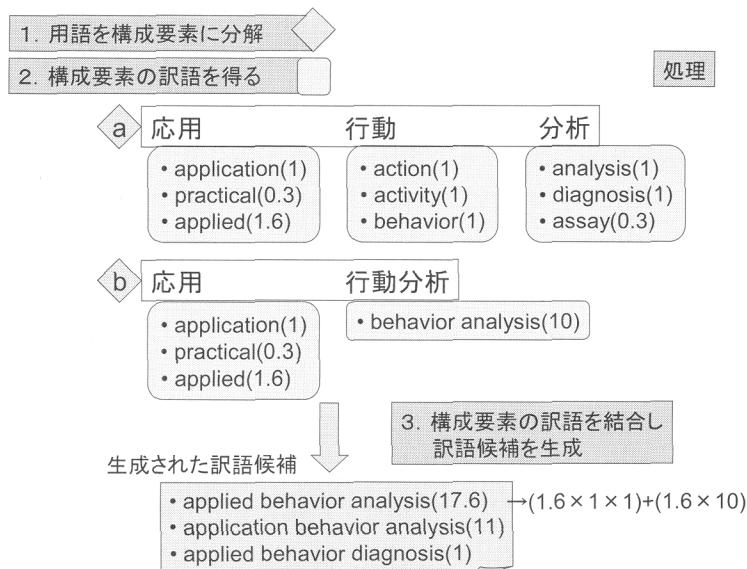


図 2 日本語の専門用語「応用行動分析」の要素合成法による訳語推定

語の専門用語「応用行動分析」を構成要素に分割する³。この例の場合、構成要素分割の結果は、図2に“a”で示した“応用”，“行動”，“分析”という分割及び，“b”で示した“応用”，“行動分析”という分割の2種類になる。次に、それぞれの構成要素を英語に翻訳する。それぞれの構成要素の訳語には、ある信頼度のスコアが与えられる。そして、“a”，“b”それぞれの分割に対し、それらの構成要素の訳語を結合することによって訳語候補を生成する。この例では、構成要素に与えられているスコアの乗算で訳語候補のスコアを計算する。“applied behavior analysis”のように分割“a”，“b”で同じ訳語候補が生成される場合には、それぞれの分割でスコアを計算し両者のスコアを加算するものとした。分割“a”では，“applied”と“behavior”と“analysis”を結合して“applied behavior analysis”が生成され、スコアは、 $1.6 \times 1 \times 1 = 1.6$ となる。また、分割“b”では，“applied”と“behavior analysis”を結合して“applied behavior analysis”が生成され、スコアは、 $1.6 \times 10 = 16$ となる。そして、最終的に“applied behavior analysis”的スコアは、 $1.6 + 16 = 17.6$ と計算される。

3.2 部分対応対訳辞書の作成

専門用語の訳語推定をするためには、既存の対訳辞書の訳語情報だけでは不十分である。複合語中の単語はどのように訳されるのが自然かという情報が重要となる。そこで、複合語中の単語の訳し方を、既存の対訳辞書の複合語エントリから収集することを試みる。一般に対訳辞書のエントリは見出し語と1つ以上の訳語から構成される。このエントリを開き、見出し語と訳語を一対一の語の組にしたもの（本論文では訳語対と呼ぶ。本節では、（藤井, 石川 2000）の語基辞書の作成方法を参考にして、既存の対訳辞書（英辞郎）の複合語の訳語対から、英語及び日本語の用語の構成要素の訳語対応を推定し、このような訳語対を集めて新たな対訳辞書を作成する方法について述べる。本論文では、この、既存の訳語対の構成要素を利用して作成された対訳辞書を部分対応対訳辞書と呼ぶ。既存の対訳辞書を部分対応対訳辞書で補う方法を、図3の例を用いて説明する。既存の対訳辞書に“applied: 応用”という訳語対自体は含まれないが、1番目の英単語が“applied”かつ1番目の日本語単語が「応用」であるような複合語の訳語対が数多く含まれていると仮定する⁴。このようなとき、それらの訳語対を対応付け、構成要素の訳語対応“applied: 応用”を推定する。

より詳細には、既存の対訳辞書から、まず、日本語及び英語の用語がそれ2つの構成要素からなる訳語対を抽出し、これを別の対訳辞書 P_2 とする。次に、 P_2 中の訳語対から英語及び日本語の第一構成要素だけを抜き出して訳語対とし、これを集めて構成した対訳辞書を前方

³ ここで、既存の対訳辞書として、「英辞郎」Ver.79 (<http://www.eijiro.jp/>) と英辞郎の訳語対から作成した部分対応対訳辞書（詳細は3.2節で述べる）を用いる。

⁴ 日本語のエントリは、形態素解析器 JUMAN (<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>) で形態素列に分解されているものとする。

applied	mathematics	:	応用	数学
applied	science	:	応用	科学
applied	robot	:	応用	ロボット
		:		頻度
		↓		↓
applied		:	応用	: 40

図 3 構成要素の訳語対の推定の例（前方一致）

一致部分対応対訳辞書 B_P と呼ぶ。同様に、 P_2 中の訳語対から英語及び日本語の第二構成要素だけを抜き出して訳語対とし、これを集めて構成した対訳辞書を後方一致部分対応対訳辞書 B_S と呼ぶ。本論文では、部分対応対訳辞書 B_P と B_S に、以下の 2 つの制約を課す。

- 前方一致部分対応対訳辞書 B_P は、用語の先頭および中間位置の構成要素の訳語を得る場合にのみ参照することとし、用語の最後尾の構成要素の訳語を得るために参照することはできない。
- 後方一致部分対応対訳辞書 B_S は、用語の中間位置および最後尾の構成要素の訳語を得る場合にのみ参照することとし、用語の先頭の構成要素の訳語を得るために参照することはできない。

これらの制約は、不適切な訳語候補が生成されるのを防ぐために課した。なお、(藤井、石川 2000)においては、 B_P と B_S を統合した部分対応対訳辞書（以下、本論文では、この辞書を部分対応対訳辞書 B と呼ぶ）を作成しており、上記のような制約を課していない。 B_P と B_S を統合すると訳語対の数が増える利点はあるが、例えば、“システム応用”という用語の訳語推定を行うときに、第二構成要素である“応用”的訳語を得るために、 B_P に含まれる訳語対〈“応用”，“applied”〉が参照されるなど、過剰に参照される恐れがある。

そこで、実際に、部分対応対訳辞書に B_P 及び B_S を利用する場合と、 B を利用する場合の比較を行った。まず、英辞郎と部分対応対訳辞書を用いて、与えられた用語に対して正解訳語が生成できるかどうかの評価を行った。詳細は 4.1 節で述べるが、部分対応対訳辞書として B を用いた場合、 B_P 及び B_S を用いた場合に比べて、正解訳語が生成できる用語の割合は 2% 程度しか上回らなかった。また、訳語推定の性能においては⁵、英日方向では再現率は B を用いる場合の方が 1% 程度高く、逆に精度は B_P, B_S を用いる場合の方が 1% 弱高かった。一方、日英方向では、 B_P, B_S を用いる場合の方が、再現率は 1% 弱高く、精度は数% 高かった。この結果を総合的に判断して、本論文では、部分対応対訳辞書として B_P, B_S を用いることとした。表

⁵ 詳細は 4.2 節で述べるが、訳語候補のスコア付けの方法としては、ウェブから収集した専門分野コーパスを利用する方法の中では総合的に最も性能のよかった‘DF-CO’というスコア関数を用いた。

表 1 対訳辞書の見出し語数と訳語対数

対訳辞書	見出し語数		訳語対の個数
	英語	日本語	
英辞郎	1,292,117	1,228,750	1,671,230
P_2	217,861	186,823	235,979
B_P	37,090	34,048	95,568
B_S	20,315	19,345	62,419
B	48,000	42,796	147,848

- 英辞郎 : 既存の汎用対訳辞書 (Ver.79)
 P_2 : 両言語とも 2 構成要素からなる英辞郎の訳語対の集合
 B_P : 前方一致部分対応対訳辞書 B_P
 B_S : 後方一致部分対応対訳辞書 B_S
 B : 部分対応対訳辞書 B

1 に、英辞郎、対訳辞書 P_2 、および、部分対応対訳辞書 B_P, B_S, B における、見出し語の数⁶と訳語対の個数を示す。

3.3 訳語候補の生成

本節及び次節において、3.2 節で準備した対訳辞書及び、2.2 節で述べた専門分野コーパスまたはウェブ全体を利用して、与えられた専門用語の訳語推定を要素合成法によって行う方法の詳細を述べる。本節では、要素合成法により訳語候補を生成する過程を定式化する。そして、次節で、本論文で実際に評価した訳語候補のスコア関数の詳細について述べる。

まず、 y_S を訳語推定すべき専門用語とする。ここで、 S が英語であれば w_i を単語、 S が日本語であれば w_i を形態素として、 y_S は以下のように w_i の列で表される。

$$y_S = w_1, w_2, \dots, w_m \quad (1)$$

例えば、 y_S が “応用行動分析” であれば、 $w_1 = “応用”$ 、 $w_2 = “行動”$ 、 $w_3 = “分析”$ となる。要素合成法では、対訳辞書中の訳語対の見出し語と照合する用語は、一個以上の単語もしくは形態素から構成されると考える。そして、 y_S を一個以上の単語もしくは形態素から構成される単位に分割し、各単位の訳語を結合することにより訳語候補を生成する。以下、まず y_S を上記の単位 s_j の列に分割する。

$$y_S = w_1, w_2, \dots, w_m \equiv s_1, s_2, \dots, s_n \quad (2)$$

⁶ 英辞郎は英日辞書であるので、本来は日本語の見出し語は存在しない。本論文では英辞郎を編集することによって日英版を作成したので、表 1 にはその見出し語数を掲載している。

ただし、各 s_j は一個以上の w_i の列を表す。例えば、 y_S を“応用 行動 分析”とすると、以下の 3通りの分割が考えられる⁷。

$$\begin{aligned} s_1 &= \text{“応用”, } s_2 = \text{“行動 分析”} \\ s_1 &= \text{“応用”, } s_2 = \text{“行動”, } s_3 = \text{“分析”} \\ s_1 &= \text{“応用 行動”, } s_2 = \text{“分析”} \end{aligned} \quad (3)$$

また、 y_S を“applied behavior analysis”とすると、以下の 3通りの分割が考えられる

$$\begin{aligned} s_1 &= \text{“applied”, } s_2 = \text{“behavior analysis”} \\ s_1 &= \text{“applied”, } s_2 = \text{“behavior”, } s_3 = \text{“analysis”} \\ s_1 &= \text{“applied behavior”, } s_2 = \text{“analysis”} \end{aligned} \quad (4)$$

次に、対訳辞書から得られた s_i の訳語を t_i とすると、 y_S の訳語候補 y_T は、以下のように y_S と同じ語順で構成される⁸。

$$y_T = t_1, t_2, \dots, t_n \quad (5)$$

そして、訳語候補 y_T にスコアを与えることを考える。先行研究と同様に、本論文においても、対訳辞書を用いて y_S と y_T の対応の適切さを推定し、スコアを与える（これを対訳辞書スコアと呼ぶ。）。ただし、 y_T 全体の対訳辞書スコアは、訳語対 $\langle s_i, t_i \rangle$ のスコア $q(\langle s_i, t_i \rangle)$ の積で構成される。また、それとは別に、目的言語コーパス内で訳語候補 y_T がどの程度出現するかによって y_T の適切さを評価し、スコアを与える（これをコーパススコア $Q_{corpus}(y_T)$ と呼ぶ。）。訳語候補 y_T のスコアは、この 2つのスコアの積により構成されるとする。

$$\prod_{i=1}^n q(\langle s_i, t_i \rangle) \cdot Q_{corpus}(y_T) \quad (6)$$

（ただし、訳語対のスコア q とコーパススコア Q_{corpus} の詳細は 3.4 節で述べる。）

⁷ 4.1 節で導入する評価用用語集合では、訳語推定すべき用語 y_S 全体が英辞郎に含まれる用語は除外している。これを除くと、この例の場合、3通りの分割が考えられる。

⁸ 英語の専門用語の中には “angle of radiation” のように前置詞を含むものがある。この用語の日本語訳語は「放射角」であるが、英語用語と日本語用語の間に双方向に適切な訳語候補として生成できるようにするためには、“of” の前後の語順の入れ替えや “of” を挿入または削除する操作を考慮する必要がある。本論文では、上記の場合も含めて、以下の様な英語・日本語の用語の組において、双方向に訳語候補の生成ができるような規則を実装した。

英語の用語	日本語の用語
angle of radiation radiation angle	⇒ { 放射 角 放射 の 角 }

なお、本論文では前置詞 “of” のみに関してこの規則を実装した。また、〈光ファイバーケーブル, optical-fiber cable〉のように、英語または日本語の用語どちらかにのみ、ハイフン記号を含む場合がある。このような場合に双方向に訳語候補生成を行うためには、ハイフンの挿入及び削除を考慮する必要がある。前置詞を含む場合と同様に、ハイフンが挿入または削除される可能性を考慮した訳語候補生成を行う規則を実装した。

実際には、例(3), (4)で示したように、 y_S には複数の分割の仕方が考えられるので、本論文ではそれぞれの分割の仕方に対して式(6)によりスコアを計算し、それらの和を訳語候補 y_T のスコアとする。

$$Q(y_S, y_T) = \sum_{y_S=s_1, s_2, \dots, s_n} \prod_{i=1}^n q(\langle s_i, t_i \rangle) \cdot Q_{corpus}(y_T) \quad (7)$$

例えば、 y_S = “応用 行動 分析”， y_T = “applied behavior analysis” の場合を考える。訳語対 $\langle y_S, y_T \rangle$ が既存の対訳辞書に含まれず、かつ、訳語対〈“応用”, “applied”〉, 〈“行動”, “behavior”〉, 〈“分析”, “analysis”〉, 〈“行動分析”, “behavior analysis”〉が既存の対訳辞書に含まれるとき、 y_T を生成することができる y_S の分割を、 y_T の生成に用いる訳語対と共に以下に示す。

- s_1 = “応用”, s_2 = “行動”, s_3 = “分析”
 - 〈“応用”, “applied”〉, 〈“行動”, “behavior”〉,
 - 〈“分析”, “analysis”〉
- s_1 = “応用”, s_2 = “行動 分析”
 - 〈“応用”, “applied”〉, 〈“行動 分析”, “behavior analysis”〉

$\langle y_S, y_T \rangle$ のスコアは、上記の2通りの分割に対して、それぞれ対訳辞書スコアとコーパススコアの積を求めたものの和となる。

次に、訳語候補生成の方法を説明する。単語または形態素数の多い用語の訳語推定を行う場合、単語または形態素の訳語のすべての組み合せを生成すると、計算機のメモリ消費量が指数関数的に増えてしまう。そこで、本論文では、この問題を避けるために、動的計画法のアルゴリズムを採用し、訳語候補の生成と枝刈りを行う。

式(1)で、訳すべき用語 y_S を以下のように単語または形態素の列で定義した。

$$y_S = w_1, w_2, \dots, w_m$$

ここで、単語または形態素 w_i の区切りの位置に、位置を表すラベル $0, \dots, m$ を付与する。

$$y_S = _0 w_1 _1 w_2 _2 \cdots _{m-1} w_m _m \quad (8)$$

この位置を表すラベルを利用して、位置 i と位置 j の間の単語または形態素の列を表す記号 $w_{j,k}$ を導入する。

$$w_{j,k} \equiv w_{j+1}, w_{j+2}, \dots, w_k \quad (9)$$

ただし、 $w_{0,0} \equiv \varepsilon$ とする。ここで、 ε は空文字を表すものとし、 y を1つ以上の単語または形態素の列とすると $\varepsilon y = y$ とする。

まず、動的計画法による y_S の訳語推定の概略を述べる。先頭から k 番目までの単語または形態

素の列 $w_{0,k}$ に対して生成された訳語候補の集合を $Tran(w_{0,k})$ とすると, $y_S = w_{0,m} = w_1, \dots, w_m$ の訳語を得るには $Tran(y_S = w_{0,m})$ からスコア 1 位の訳語候補を取り出せばよい. ここで, 各 $Tran(w_{0,k})$ ($k = 1, \dots, m$) は以下の式に従って, 再帰的に計算される.

$$Tran(w_{0,k}) = top(merge(\bigcup_{i=0}^{k-1} concat(Tran(w_{0,i}), tran(w_{i,k}))), r) \quad (10)$$

この式では, $w_{0,k} = w_1, \dots, w_k$ をある位置 i で $w_{0,i} = w_1, \dots, w_i$ と $w_{i,k} = w_{i+1}, \dots, w_k$ の 2 つに分割する. 分割の場所 i は先頭 $i = 0$ から順に $i = k - 1$ まで移動させていく. それぞれの分割の仕方において, $w_{0,i}$ に対しては再帰的に訳語候補の集合 $Tran(w_{0,i})$ を求め, $w_{i,k}$ に対しては $w_{i,k}$ を見出し語として対訳辞書から訳語の集合 $tran(w_{i,k})$ を得る. そして, 両者を $concat$ により結合することにより, 新しい訳語候補を生成する. このとき, 同一の訳語候補が複数の異なる分割の仕方から生成される場合がある. その場合は, $merge$ により, それらの訳語候補のスコアがまとめられる. 最後に, top によりスコア上位 r 個の訳語候補のみ出力することで, 訳語候補の枝刈りを行い, この出力を $Tran(w_{0,k})$ とする.

実際に式 (10) を用いて, 図 2 の例をもとに $y_S = “応用”, “行動”, “分析”$ の訳語候補の集合 $Tran(y_S = w_{0,3})$ が生成される様子を説明する. ただし, ここでは枝刈り後出力される訳語候補数 r を 3 とする. 式 (10) の $i = 0, \dots, k - 1$ のループに注目すると, 以下のように訳語候補が生成されていくことがわかる.

($i = 0$) $w_{0,3}$ は $w_{0,0} = \varepsilon$ と $w_{0,3} = “応用”, “行動”, “分析”$ に分割される. $w_{0,3} = “応用”, “行動”, “分析”$ は対訳辞書に訳語がないので, $concat$ の出力は空集合となる.

($i = 1$) $w_{0,3}$ は $w_{0,1} = “応用”$ と $w_{1,3} = “行動”, “分析”$ に分割される. $Tran(w_{0,1})$ により, $w_{0,1}$ の訳語候補の集合を再帰的に求め, $tran(w_{1,3})$ により, 対訳辞書から得られる $w_{1,3}$ の訳語の集合を求める. そして, それらの訳語候補を $concat$ により結合し訳語候補を生成する. $Tran(w_{0,1}) = \{“application”, “practical”, “applied”\}$, $tran(w_{1,3}) = \{“behavior analysis”\}$ のとき生成される訳語候補を以下に示す.

- “application behavior analysis”
- “practical behavior analysis”
- “applied behavior analysis”

($i = 2$) $w_{0,3}$ は $w_{0,2} = “応用”, “行動”$ と $w_{2,3} = “分析”$ に分割される. $Tran(w_{0,2})$ により, $w_{0,2}$ の訳語候補の集合を再帰的に求め, $tran(w_{2,3})$ により, 対訳辞書から得られる $w_{2,3}$ の訳語の集合を求める. そして, それらの訳語候補を $concat$ により結合し訳語候補を生成する. $Tran(w_{0,2}) = \{“applied action”, “applied activity”, “applied behavior”\}$, $tran(w_{2,3}) = \{“analysis”, “diagnosis”, “assay”\}$ のとき生成される訳語候補を以下に示す.

- “applied action analysis”
- “applied action diagnosis”
- “applied action assay”
- “applied activity analysis”
- “applied activity diagnosis”
- “applied activity assay”
- “applied behavior anaylysis”
- “applied behavior diagnosis”
- “applied behavior assay”

以上の操作が終したら、 y_S に対して複数個の訳語候補が生成された状態となる。生成された訳語候補に同じものが存在した場合、関数 $merge$ によりこれらがまとめられ、最後に関数 top によりスコア上位 r 個の訳語候補が出力される。

3.4 訳語候補のスコア付け

3.4.1 対訳辞書スコア

訳語推定対象の用語 y_S と訳語候補 y_T の対応の適切さを対訳辞書を用いて測定するための「対訳辞書スコア」を 3.3 節で導入した。この対訳辞書スコアは、訳語候補 y_T を生成するときに使用した訳語対 $\langle s_i, t_i \rangle$ のそれぞれの適切さを関数 q により測定し、それらの積で計算されるものであった。本節では、対訳辞書に基づいて訳語対 $\langle s_i, t_i \rangle$ のスコアを計算するための関数 q を 2 種類定義する。

頻度-長さ (DF)

“natural language processing” という用語が、既存の対訳辞書に含まれないため、訳語推定の対象となる場合を考える。 \langle “natural”, “自然な” \rangle , \langle “language”, “言語” \rangle , \langle “processing”, “処理” \rangle の 3 つの訳語対から生成される“自然な言語処理”という訳語候補よりも、 \langle “natural language”, “自然言語” \rangle のような単語数または形態素数の多い訳語対と \langle “processing”, “処理” \rangle を利用して得られる訳語候補“自然言語処理”の方が信頼度が高いと思われる。また、表 1 の部分対応対訳辞書に含まれる訳語対は、英辞郎に含まれる複合語の訳語対から、英語及び日本語の構成要素の訳語対応を推定することにより作成された訳語対であるため、対訳辞書 P_2 に出現する頻度の少ない訳語対よりも、出現する頻度の多い訳語対の方が信頼度が高いと思われる。以上のような、訳語対の長さと頻度に基づく経験的な選好に基づいて、訳語対を順位付けする方法について述べる。

まず、スコア付けの対象となる対訳辞書の訳語対は、以下のように分類できる。

- 英辞郎の訳語対（利用できる情報：単語数または形態素数）

- 単語数または形態素数が 2 以上の訳語対 ((a) とする)
- 1 単語または 1 形態素の訳語対 ((b) とする)
- 部分対応対訳辞書の訳語対 (利用できる情報: 対訳辞書 P_2 に出現する頻度) ((c) とする)

ここではスコア付けの方針を決める問題を、上記の (a), (b), (c) で示した 3 種類の訳語対の間に優先順位を付けることに帰着させて考える。 (a), (b), (c) の優先順位として、本論文では、まず、(a) の訳語対に与えるスコアを極めて高く設定し、(b) または (c) の訳語対のスコアを必ず上回るようにする。次に、(b) と (c) の訳語対の間のスコアの大小関係については、(c) の訳語対が対訳辞書 P_2 に出現する頻度に閾値を設け、(c) の訳語対の頻度が頻度閾値と同じであれば、(b) の訳語対のスコアと同じにし、(c) の訳語対の頻度が頻度閾値より大きければ、(b) の訳語対のスコアより大きくし、そして、(c) の訳語対の頻度が頻度閾値より小さければ、(b) の訳語対のスコアより小さくする。本論文では、この頻度閾値を 10 に設定した。この頻度閾値を変化させることにより、英辞郎に含まれる 1 単語または 1 形態素の訳語対のスコアと、部分対応対訳辞書に含まれる訳語対のスコアの大小関係が変化するため、訳語推定の性能にもある程度影響を与える。しかしながら、本論文の目的は、コーパスとしてウェブ全体を用いる方法と、ウェブから収集した専門分野コーパスを利用する方法の比較にあるので、最適なパラメータの値の追求は行わなかった。

この優先順位を実現するため、英辞郎の訳語対のスコアには単語数または形態素数を指数とする関数を用い、部分対応対訳辞書の訳語対のスコアには頻度の対数を用いることで、単語数または形態素数が 2 以上の英辞郎の訳語対のスコアが、部分対応対訳辞書の訳語対のスコアよりも大きくなるようにした。訳語対 $\langle s, t \rangle$ のスコア $q(\langle s, t \rangle)$ の定義として、以下の式を採用する。

$$q(\langle s, t \rangle) = \begin{cases} 10^{(compo(s)-1)} & (\langle s, t \rangle \text{ in 英辞郎}) \\ \log_{10} f_p(\langle s, t \rangle) & (\langle s, t \rangle \text{ in } B_P) \\ \log_{10} f_s(\langle s, t \rangle) & (\langle s, t \rangle \text{ in } B_S) \end{cases} \quad (11)$$

ここで、 $compo(s)$ は s の単語または形態素の数を表すものとし、 $f_p(\langle s, t \rangle)$ は、 P_2 中に第一要素として $\langle s, t \rangle$ が出現する回数を表すものとし、 $f_s(\langle s, t \rangle)$ は、 P_2 中に第二要素として $\langle s, t \rangle$ が出現する回数を表すものとする。式 (11) では、対数関数の底の値が部分対応対訳辞書の訳語対の頻度閾値に対応する。すなわち、部分対応対訳辞書の訳語対で対訳辞書 P_2 に 10 回出現する訳語対と、英辞郎に含まれる 1 単語または 1 形態素の訳語対のスコアが等しくなる。なお、このスコアでは、部分対応対訳辞書に一度しか現れない訳語対のスコアはゼロとなる。この場合、訳語として利用しないものとする。

式 (11) に示した訳語対のスコア関数の積で定義される対訳辞書スコアを、以下では DF と呼ぶものとする。

確率 (DP)

(藤井, 石川 2000) は, 対訳辞書に基づく y_S と y_T の対応の適切さを, 確率 $P(y_S|y_T)$ を計算することにより評価した. このスコアは, 条件付き確率 $P(s_i|t_i)$ の積で定義される. (藤井, 石川 2000) は対訳辞書として部分対応対訳辞書 B のみを用いているため, 同じ設定とするには, 本論文でも部分対応対訳辞書 B のみを用いなければならない. しかしながら, 部分対応対訳辞書 B のみを用いた実験を行った結果, 英辞郎と部分対応対訳辞書 B を併用する場合に比べ, 訳語推定の性能 (精度・再現率) が 10% 前後も低いことがわかった. このため, 本論文では, 部分対応対訳辞書 B に加え英辞郎も用いて, 条件付き確率 $P(s_i|t_i)$ に基づく対訳辞書スコアを評価することとした. 本論文では, 英辞郎と部分対応対訳辞書 B を併用できるようにするために, 以下の式に示す拡張を行った.

$$q(\langle s, t \rangle) = P(s|t) = \frac{f_{prob}(\langle s, t \rangle)}{\sum_{s_j} f_{prob}(\langle s_j, t \rangle)} \quad (12)$$

$$f_{prob}(\langle s_j, t \rangle) = \begin{cases} 10 & (\langle s, t \rangle \text{ in 英辞郎}) \\ f_B(\langle s_j, t \rangle) & (\langle s, t \rangle \text{ in } B) \end{cases} \quad (13)$$

上式では, 英辞郎の訳語対の頻度は 10 とみなすものとした⁹.

式 (12), (13) に示した訳語対のスコア関数から計算される対訳辞書スコアを以下では DP と呼ぶものとする.

3.4.2 コーパスに基づくスコア

訳語候補 y_T の適切さを目的言語コーパスを用いて測定するための「コーパススコア」を 3.3 節で導入した. 本論文では, コーパススコアとして以下に示す 3 種類を評価した.

- 頻度 (CF): 目的言語コーパスにおける訳語候補 y_T の生起頻度

$$Q_{corpus}(y_T) = freq(y_T) \quad (14)$$

- 確率 (CP): 以下のバイグラムモデルによって推定される, 訳語候補 y_T の生起確率. (藤井, 石川 2000) で用いられたコーパススコアの評価を目的とする. 本来は t_i を単語または形態素とすべきであるが, 実装の都合上, t_i を対訳辞書から得られた訳語とする. したがって, t_i は 1 つ以上の単語または形態素から構成される¹⁰.

⁹ 辞書スコア ‘DF’ では, 頻度 10 の部分対応対訳辞書の訳語対のスコアと, 構成要素長が 1 の用語の英辞郎の訳語のスコアと同じにしている. これに合わせるため, 英辞郎の訳語対の頻度を 10 とみなすものとした.

¹⁰ 前節述べたように, (藤井, 石川 2000) のスコア関数の評価に際しては, 対訳辞書スコアにおいて, 部分対応対訳辞書 B と英辞郎を併用している. ここで, 英辞郎の訳語には複数の単語または形態素で構成されるものがあるが, このような場合, 厳密には, 訳語を単語または形態素に分割して, 単語または形態素のバイグラムに基づいて式 (15) の計算をしなければならない. しかしながら, 実装上の手間を避けるため, ここでは, 対訳辞書から得られた訳語そのまま用い, t_i は 1 つ以上の単語または形態素から構成されたとした.

$$Q_{corpus}(y_T) = P(t_1) \cdot \prod_{i=1}^{n-1} P(t_{i+1}|t_i) \quad (15)$$

- 生起 (CO): 目的言語コーパスに訳語候補 y_T が生起するかどうか

$$Q_{corpus}(y_T) = \begin{cases} 1 & y_T \text{ がコーパス中に生起する} \\ 0 & y_T \text{ がコーパス中に生起しない} \end{cases} \quad (16)$$

3.4.3 スコア関数

表 2 に示すように、本論文では、辞書に基づくスコアとコーパスに基づくスコアに対して、12種類の組み合せのスコア関数を作成し評価を行った。この表において、「p(prune)」は、動的計画法のアルゴリズムを用いた訳語候補生成の過程において、式(10)の *top* を実行することで、生成された訳語候補の部分列の順位付けと枝刈りにそのスコアが用いられる事を示す。「f(final)」は、生成された訳語候補の最終結果の順位付けにそのスコアが用いられる事を示す。また、列「コーパス」において、「専門分野コーパス」は、あらかじめウェブから専門分野コーパスを収集し、その後、このコーパスを用いて生成された訳語候補の検証を行うことを示す。「ウェブ全体」は、サーチエンジンを通してウェブ全体を利用して訳語候補の検証を行うことを示す。

スコア関数の命名方法は、「対訳辞書スコア名-コーパススコア名」の原則に基づく。例えば、

表 2 訳語候補のスコア関数と構成要素

スコア関数	対訳辞書スコア		コーパススコア			コーパス	
	頻度-長さ (DF)	確率 (DP)	頻度 (CF)	確率 (CP)	生起 (CO)	専門分野 コーパス	ウェブ 全体
DF-CF	p/f		p/f			o	
DF-CF _f	p/f		f			o	
DF-CP	p/f			p/f		o	
DF-CO	p/f				p/f	o	
DF-CO _f	p/f				f	o	
DP-CF		p/f	p/f			o	
DP-CP		p/f		p/f		o	
CF			p/f			o	
CP				p/f		o	
DF-CF _{f-w}	p/f		f				o
DF-CO _{f-w}	p/f				f		o
DF	p/f						

p(prune):枝刈りに利用, f(final):最終スコアに利用

スコア関数‘DF-CO’は、対訳辞書スコアに‘DF’を用い、コーパススコアに‘CO’を用いたスコア関数である。ここで、式(10)の*top*による訳語候補の枝刈りについて考えると、不要な候補を早い段階で削減するため、基本的には対訳辞書スコアとコーパススコアの両方を用いるべきである。しかしながら、コーパスとしてウェブ全体を用いる場合は、サーチエンジンの検索に要する時間を考慮すると、訳語候補の生成過程でコーパススコアを利用するることは効率的ではない。そこで、訳語候補の枝刈りにはコーパススコアを用いず、訳語候補の最終的なスコア計算のみにコーパススコアを用いる。コーパススコアを枝刈りに用いない場合は、‘DF-COf’の様に、コーパススコア名の後ろに‘f’を付加する。そして、コーパスとしてウェブ全体を用いる場合は、‘DF-COf-w’の様に、‘-w’を付加する。

本論文で評価したスコア関数は、コーパススコアの計算において用いるコーパスの違いにより、ウェブから収集した専門分野コーパスを用いるタイプ、サーチエンジンを通してウェブ全体を用いるタイプ、コーパスを一切用いないタイプの、3つのタイプに分けることができる。対訳辞書スコアには、訳語対が部分対応対訳辞書に出現する頻度と訳語対の構成要素長に基づく‘DF’と、条件付き確率 $P(s|t)$ に基づく‘DP’の2つがある。コーパススコアには、訳語候補がコーパスに生起する頻度に基づく‘CF’、訳語候補がコーパスに生起する確率に基づく‘CP’、訳語候補がコーパスに生起するか否かに基づく‘CO’の3つがある。ここで、4.2節で示す実験結果においては、対訳辞書‘DF’を用いたスコア関数と‘DP’を用いたスコア関数の間で性能に大きな差はないが、‘DF’を用いた方が若干精度が高かった。そこで本論文では、精度を重視する立場に立ち、対訳辞書スコアとして主に‘DF’を用いて評価を行う¹¹。

以下、本論文で実際に評価した辞書スコアとコーパススコアの組み合わせについて説明する。

コーパスとしてウェブから収集した専門分野コーパスを用いる場合には、対訳辞書スコア‘DF’を用いたスコア関数とコーパススコアとの組み合わせでは、‘CF’, ‘CP’, ‘CO’の3種類を網羅したが、それらの性能に大差はなかった。そこで、対訳辞書スコア‘DP’では、大きく性質の異なるコーパススコアである‘CF’, ‘CP’との組み合わせを評価した。ここで、スコア関数‘DP-CP’は、(藤井, 石川 2000)で提案されたモデルに、部分対応対訳辞書に加え英辞郎自体も用いることができるよう拡張を加えたスコア関数である。

一方、コーパスとしてウェブ全体を用いる場合は、辞書スコアとしては‘DF’を用いた。また、コーパススコア‘CP’は、‘CF’や‘CO’と比べ、サーチエンジンの検索回数が多くなるので、評価の対象から除外した。さらに、上述したように、サーチエンジンの検索時間の都合で、コーパススコアによる枝刈りは行わない。以上をまとめると、コーパスとしてウェブ全体を用いるスコア関数としては、‘DF-CFf-w’と‘DF-COf-w’の2種類を評価する。そして、この2つのス

¹¹ 本論文の焦点は、ウェブから収集した専門分野コーパスを用いる方法と、サーチエンジンを通してウェブ全体を用いる方法の比較にある。従って、定義し得るスコア関数を網羅的に評価することは行っていない。

コア関数との直接的な比較のため、ウェブから収集した専門分野コーパスを用いるスコア関数として、コーパススコアによる枝刈りを行わない DF-CF_f と DF-CO_f についても評価を行う。

最後に、対訳辞書スコアまたはコーパススコアどちらかのみを用いるスコア関数の評価のために、次のスコア関数を評価する。辞書スコアのみで訳語候補のスコア付けをした場合の評価のため、スコア関数 ‘DF’ を評価する¹²。コーパススコアのみで訳語候補のスコア付けをした場合の評価のため、スコア関数 ‘CF’ 及び ‘CP’ を評価する¹³。

4 実験と評価

4.1 評価用用語集合

実験では、図 1 で示したように、既存の対訳辞書に含まれている用語と含まれていない用語が混在した形で複数の専門用語が与えられるものとし、既存の対訳辞書に載っていない用語の訳語推定の評価を行う。本論文では、言語 $\langle S, T \rangle$ の組を〈英語、日本語〉または、〈日本語、英語〉とする。評価セットを作成するため、まず、表 3 に示す既存の 4 種類の日英専門用語対訳辞書「マグローヒル科学技術用語大辞典」(マグローヒル科学技術用語大辞典編集委員会 1998)，

表 3 評価用用語の数

辞書	分野	Y _S	S = 英語		S = 日本語	
			X _S ^U	MBytes	X _S ^U	MBytes
マグローヒル 科学技術用語 大辞典	電磁気学	30	36	28	32	99
	電気工学	41	34	21	25	71
	光学	31	42	37	22	48
岩波 情報科学辞典	プログラム言語	28	37	34	38	135
	プログラミング	28	29	33	29	110
英和コンピュータ 用語大辞典	(コンピュータ)	99	91	67	69	232
25 万語 医学用語大辞典	解剖学	91	91	73	33	66
	疾患	86	91	83	53	100
	化学物質及び薬物	84	94	54	74	131
	物理化学及び統計学	99	88	56	58	135
合計		617	633	482	433	1127

¹² 4.2 節の評価実験では、スコア関数 ‘DF’ は極めて低い F 値であった。本論文ではスコア関数 ‘DP’ は評価していないが、同様の傾向であると思われる。

¹³ スコア関数 ‘CO’ は、辞書スコアは利用せずコーパススコア ‘CO’ のみを用いるスコア関数であるが、訳語候補のスコアが 0 か 1 となってしまい、順位付けできないので取り扱わない。

「岩波情報科学辞典」(長尾, 石田, 稲垣, 田中, 辻井, 所, 中田, 米澤 1990), 「英和コンピュータ用語大辞典」(コンピュータ用語辞典編集委員会 2001), 「25万語医学用語大辞典」(医学用電子化AI辞書研究会 1996)の10分野から, 以下の条件を満たす2種類の訳語対集合を無作為に選定した。1種類目は, 英辞郎にも訳語対として存在し, かつ英語用語・日本語用語共にヒット数が100以上の訳語対であり, この種類の訳語対の集合を既知訳語対集合 X_{ST} とする。2種類目は, 以下の条件を満たす訳語対 $\langle t_e, t_j \rangle$ の集合であり, これを未知訳語対集合 Y_{ST} とする。ただし, t_e は英語の用語, t_j は日本語の用語を表すものとする。

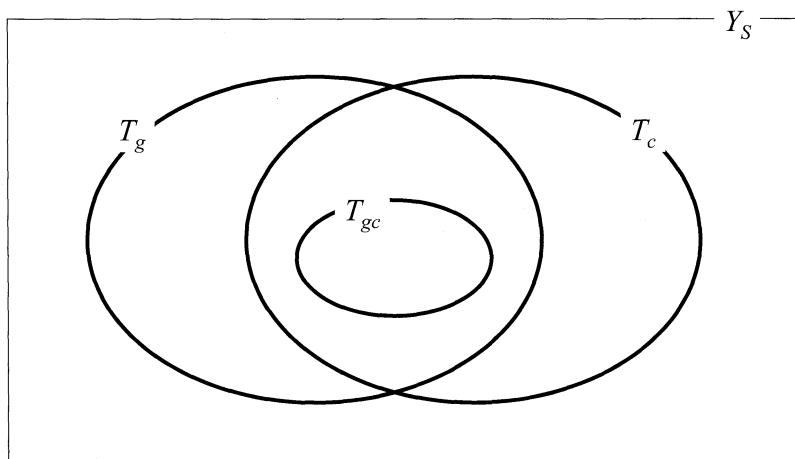
- t_e, t_j 共に英辞郎に見出し語として存在しない
- t_e は2語以上, t_j は2形態素以上からなる
- t_e 及び t_j のヒット数は10以上

次に, $S = \text{英語}$, $S = \text{日本語}$, それぞれの場合において, 既知訳語対集合 X_{ST} から, 英辞郎に含まれる訳語が一個の用語の集合 X_S^U を作成した。そして, 2.2節で述べた方法で, X_S^U の用語に対して, それら用語の英辞郎に含まれる訳語の集合 X_T^U を利用して, 各分野毎にウェブから専門分野コーパスを収集した。同様に, $S = \text{英語}$, $S = \text{日本語}$, それぞれの場合において, 未知訳語対集合 Y_{ST} から, 評価用用語集合 Y_S を作成した。そして, Y_S のそれぞれの用語に対して, 未知訳語対集合 Y_{ST} にもともと含まれる訳語に加えて, 必要であれば人手で一個以上の正解訳語を付与した。この結果, 正解訳語の個数の平均は, $S = \text{英語}$ のとき1.31個, $S = \text{日本語}$ のとき1.62個となった。10分野のそれぞれに対して, 表3に X_S^U 及び Y_S に含まれる用語の個数, 及び, ウェブから収集したコーパスのサイズを分野毎に示す。

続いて, Y_S 及び, それに属する用語の正解訳語の性質について述べる。 Y_S に属する用語の構成要素数の平均は, S が英語のとき2.28語(2語の用語は437個で全体の70.8%, 3語以上の用語は180個で全体の29.2%), S が日本語のとき2.47形態素(2形態素の用語は396個で全体の64.2%, 3形態素以上の用語は221個で全体の35.8%)であった。また, Y_S の用語とその正解訳語が構成的に対応しているかどうかを調べると, S が英語のとき90.6%, S が日本語のとき92.5%であった¹⁴。

次に, 集合 Y_S の用語のどの程度が訳語推定可能かを調べるために, Y_S の部分集合として, 図4に示す T_g , T_c , T_{gc} を定義する。 T_g は, 対訳辞書として英辞郎と部分対応対訳辞書 B_P , B_S を利用して, 3.3節で述べた方法で, 出力される訳語候補数 r を無限大にしたときに, 正解訳語を生成可能な用語の集合である。 T_c は, その用語が属する分野の専門分野コーパスに正解訳語が含まれている用語の集合である。 T_{gc} は, 図4に示したように, 生成可能かつ専門分野コーパ

¹⁴ 1章では, 未知訳語対集合 Y_{ST} に含まれる訳語対が構成的かどうかを調査した結果を述べている。これに対して, 本節では, 未知訳語対集合 Y_{ST} に含まれる訳語対に対して, 新たに人手で正解訳語を追加したものに対して構成的かどうか評価を行っているので, 1章の結果よりも割合が増加し, また, S が英語のときと日本語のときで割合も異なる。



$t \in Y_S, S(t)$ は t の正解訳語の集合

$$T_g = \{t | \exists s \in S(t), s \text{ は生成可能}\}$$

$$T_c = \{t | \exists s \in S(t), s \text{ はコーパスに存在}\}$$

$$T_{gc} = \{t | \exists s \in S(t), s \text{ は生成可能かつ } s \text{ はコーパスに存在}\}$$

図 4 正解訳語の生成可能性／コーパス中の出現による集合 Y_S の部分集合の分類

スに含まれる正解訳語が存在するという条件を満たす用語の集合である。言い換えると、英辞郎、および部分対応対訳辞書 B_P, B_S を用いた訳語候補生成手法において、専門分野コーパスを利用した場合には、 T_{gc} に属する用語に対してのみ、正解訳語を生成できる可能性がある¹⁵。 T_g と T_c の積集合の中には、生成可能かつ専門分野コーパスに含まれる正解訳語が存在しない用語も含まれているが、このような用語に対しては正解訳語を生成することができない。 T_{gc} は、 T_g と T_c の積集合の部分集合となっていることに注意されたい。

以上の定義をふまえて、 T_g, T_c, T_{gc} に属する用語の割合を分野毎に調べた結果を表 4 に示す。これより、英辞郎と部分対応対訳辞書 B_P, B_S を利用して正解訳語が生成可能な用語の割合は、英日方向で 71.8%，日英方向で 75.7% であることがわかる。一方、英辞郎のみを利用して生成可能な用語の割合を評価すると、英日方向で 50.4%，日英方向で 56.6% であった。このことから、部分対応対訳辞書が有効であることがわかる。また、英辞郎、および部分対応対訳辞書 B_P, B_S を用いた訳語候補生成手法において、専門分野コーパスを利用した場合には、正解訳語を生成できる用語の割合の上限は、 T_{gc} 欄より、英日方向で 53.8%，日英方向で 51.9% で

¹⁵ コーパススコア ‘CP’ を用いた場合は、訳語推定対象の用語が T_g に属し、かつ、コーパススコアの確率の各項 $P(t_1), P(t_{i+1}|t_i)$ がゼロでなければ、正解訳語そのものがコーパスに存在しなくとも、正解訳語を生成できる可能性がある。

表 4 英辞郎と部分対応対訳辞書 B_P, B_S における正解訳語の生成可能性と正解訳語がコーパスに存在するかどうかの調査

分野	英語→日本語			日本語→英語		
	生成可能 (T_g)	コーパス に存在 (T_c)	生成可能かつ コーパスに存在 (T_{gc})	生成可能 (T_g)	コーパス に存在 (T_c)	生成可能かつ コーパスに存在 (T_{gc})
電磁気学	60%	93%	57%	87%	90%	73%
電気工学	71%	78%	59%	71%	68%	51%
光学	61%	65%	35%	71%	68%	52%
プログラム言語	86%	93%	79%	82%	100%	82%
プログラミング	75%	96%	71%	82%	86%	75%
コンピュータ	74%	52%	34%	80%	61%	46%
解剖学	78%	92%	74%	80%	55%	45%
疾患	69%	83%	55%	76%	70%	56%
化学物質及び薬物	54%	63%	39%	56%	54%	36%
物理化学及び統計学	85%	71%	58%	80%	61%	53%
全体	71.8%	74.9%	53.8%	75.7%	65.3%	51.9%

表 5 英辞郎と部分対応対訳辞書 B における正解訳語の生成可能性

分野	英語→日本語	日本語→英語
電磁気学	60%	87%
電気工学	80%	78%
光学	61%	71%
プログラム言語	86%	82%
プログラミング	79%	82%
コンピュータ	75%	81%
解剖学	79%	81%
疾患	71%	78%
化学物質及び薬物	55%	57%
物理化学及び統計学	86%	81%
全体	73.6%	77.0%

あることがわかる。参考として、対訳辞書として、英辞郎と部分対応対訳辞書 B を利用して、正解訳語の生成可能性を調べた結果を表 5 に示す。この結果を見ると、対訳辞書として、英辞郎と部分対応対訳辞書 B を用いる方が、正解訳語を生成可能な用語数が若干多い。しかしながら、3.2 節で述べたように、本論文では両者の性能を総合的に比較して、部分対応対訳辞書として B_P, B_S を用いている。

ここで、3.3節の脚注8で述べた、前置詞“of”及びハイフンの挿入・削除に関する規則の効果について述べる。未知語対集合 Y_{ST} の語対617個のうち、前置詞“of”を含むものは24個存在した。また、英語の用語のみにハイフンを含む語対は33個、日本語の用語のみにハイフンを含む語対は2個存在した。英日方向において、この2つの規則を加えることで、 T_g に含まれる用語の数が27個しか増加しなかった。逆に日英方向の場合、この2つの規則を加えることで、 T_g に含まれる用語の数が7個しか増加しなかった。 T_g に含まれる用語数の増加が少ないのは、正解語を人手で付与することにより、ofやハイフンを含まない正解語が追加されたためである。

4.2 スコア関数の評価

表2に示したスコア関数を用いて、集合 Y_S に対して語対推定の評価実験を行った。実験の条件として、動的計画法による語生成過程で、保持する語候補の数 r は10とした。専門分野コーパスを用いる場合は、対象の用語が属する分野の専門分野コーパスを用いる。ウェブ全体を用いる場合は、サーチエンジンとして、英日方向の場合はgooを、日英方向の場合はYahoo!を用いた。また、日英方向では、日本語の用語の分かち書きは人手で行った。

Y_S 全体に対する実験の結果を表6に示す。列‘top 1’には、スコア1位の語候補が正解である割合を、列‘top 10’には、正解語がスコア10位以内に含まれる割合を示す。ここで、 Y_S 全体に対して、スコア1位の語候補が正解である用語の割合を再現率と定義する。次に、語候補が1つ以上生成される用語に限定した評価の結果を表7に示す。表の「出力あり」の欄

表6 集合 Y_S 全体に対するスコア関数の評価（再現率）

スコア関数	英語→日本語		日本語→英語	
	top 1	top 10	top 1	top 10
DF-CF	42.5%	50.1%	41.8%	48.1%
DF-CF _f	38.2%	41.5%	39.5%	44.1%
DF-CP	43.6%	52.4%	44.6%	51.9%
DF-CO	44.7%	47.6%	43.9%	48.6%
DF-CO _f	39.9%	41.7%	39.7%	44.1%
DP-CF	46.0%	54.3%	44.7%	51.5%
DP-CP	46.7%	56.2%	48.9%	56.1%
CF	26.3%	48.0%	31.3%	43.8%
CP	25.9%	48.6%	32.6%	46.5%
DF-CF _{f-w}	52.0%	59.0%	51.1%	65.8%
DF-CO _{f-w}	44.1%	59.0%	50.1%	65.0%
DF	35.7%	59.0%	45.1%	63.2%

表 7 訳語候補が 1 つ以上生成される用語に対する評価

(a) 英語→日本語

スコア関数	出力あり	1 位正解			10 位以内正解		
		個数	精度	F 値	個数	精度	F 値
DF-CF	396	262	66.2%	51.7%	309	78.0%	61.0%
DF-CF _f	303	236	77.9%	51.3%	256	84.5%	55.7%
DF-CP	428	269	62.9%	51.5%	323	75.5%	61.8%
DF-CO	379	276	72.8%	55.4%	294	77.6%	59.0%
DF-CO _f	303	246	81.2%	53.5%	257	84.8%	55.9%
DP-CF	455	284	62.4%	53.0%	335	73.6%	62.5%
DP-CP	495	288	58.2%	51.8%	347	70.1%	62.4%
CF	456	162	35.5%	30.2%	296	64.9%	55.2%
CP	497	160	32.2%	28.7%	300	60.4%	53.9%
DF-CF _{f-w}	481	321	66.7%	58.5%	364	75.7%	66.3%
DF-CO _{f-w}	481	272	56.5%	49.5%	364	75.7%	66.3%
DF	559	220	39.4%	37.4%	364	65.1%	61.9%

(b) 日本語→英語

スコア関数	出力あり	1 位正解			10 位以内正解		
		個数	精度	F 値	個数	精度	F 値
DF-CF	372	258	69.4%	52.2%	297	79.8%	60.1%
DF-CF _f	317	244	77.0%	52.2%	272	85.8%	58.2%
DF-CP	418	275	65.8%	53.1%	320	76.6%	61.8%
DF-CO	369	271	73.4%	55.0%	300	81.3%	60.9%
DF-CO _f	317	245	77.3%	52.5%	272	85.8%	58.2%
DP-CF	428	276	64.5%	52.8%	318	74.3%	60.9%
DP-CP	489	302	61.8%	54.6%	346	70.8%	62.6%
CF	428	193	45.1%	36.9%	270	63.1%	51.7%
CP	488	201	41.2%	36.4%	287	58.8%	51.9%
DF-CF _{f-w}	522	315	60.3%	55.3%	406	77.8%	71.3%
DF-CO _{f-w}	522	309	59.2%	54.3%	401	76.8%	70.4%
DF	565	278	49.2%	47.0%	390	69.0%	66.0%

には、訳語候補が 1 つ以上生成される用語数を示した。訳語候補が 1 つ以上生成される用語に対して、スコア 1 位の訳語候補が正解である割合を精度と定義する。また、「F 値」の欄は、表 6 の値を再現率として計算した。

ウェブから収集した専門分野コーパスとウェブ全体の比較

まず、ウェブから収集した専門分野コーパスを用いる方法とウェブ全体を用いる方法の比較を行う。英日方向、日英方向で平均を取ってみると、ウェブ全体を用いるスコア関数の方が再現率が高いことがわかる。これは、表4からわかるように、集合 Y_S 全体に対して、収集した専門分野コーパスに正解訳語が含まれる割合 T_c が、英日方向で 74.9%，日英方向で 65.3% と、あまり高くないことが原因と考えられる。精度に関しては、専門分野コーパスを用いるスコア関数であれば ‘DF-COf’ の平均 79.3% が最も高く、ウェブ全体を用いるスコア関数であれば ‘DF-CF_{f-w}’ の平均 63.5% が最も高い。このことから、専門分野コーパスを用いるスコア関数の方が精度が高いことがわかる。これは、ウェブ全体には一般語や訳語推定対象の分野以外の用語が多数含まれており、不正解の訳語にも大きいスコアが与えられてしまうためと考えられる。F 値に関しては、専門分野コーパスを用いるスコア関数であれば ‘DP-CO’ の平均 55.2% が最も高く、ウェブ全体を用いるスコア関数であれば ‘DF-CF_{f-w}’ の平均 56.9% が最も高い。このことから、F 値には大きな差はないことがわかる。以上より、コーパスとしてウェブ全体を用いる手法は再現率を重視した手法と言える一方、専門分野コーパスを用いる手法は精度を重視した手法と言える。また、これらの考察から、ウェブから収集した専門分野コーパスを用いる方法において、精度を下げることなく、再現率を上げるために、対象分野の用語を十分に含み、かつ、できるだけ小さなコーパスを収集する必要があることがわかる。また、両者を相補的に統合する方法としては、まず、ウェブから収集した専門分野コーパスを用いて高い精度で訳語推定を行い、訳語候補が 1 つも得られなかった用語に対しては、ウェブ全体を用いて訳語推定を行うことが考えられる。

ウェブから収集した専門分野コーパスを用いるスコア関数の評価

次に、ウェブから収集した専門分野コーパスを用いるスコア関数を比較し評価する。

まず、最も精度が高かったスコア関数は、英日方向、日英方向とも、スコア関数 ‘DF-COf’ であった。2番目に精度が高かったスコア関数は、英日方向、日英方向とも、スコア関数 ‘DF-CF_f’ であった。‘DF-COf’ や ‘DF-CF_f’ のように、訳語候補の生成途中でコーパススコアを利用しないスコア関数が高い精度となったのは、辞書スコアのみを利用して生成される訳語候補のほとんどはコーパスに存在せず、最後にコーパスで検証することにより、これらがすべて消えてしまい、正解訳語が高いスコアとなって残った場合のみ、これが出力されるという現象による。このため、この 2 つのスコア関数の「出力あり」の個数は他のスコア関数に比べて低い値となっており、再現率は、他のスコア関数に比べ若干低い値となっている。F 値で評価した場合は、他のスコア関数と大きな差はない。F 値を下げずに、精度を上げたい場合は、このスコア関数が有効である。

次に、辞書スコア ‘DF’ を用いるスコア関数と辞書スコア ‘DP’ を用いるスコア関数の比較を

行う。スコア関数‘DF-CF’と‘DP-CF’を比較し、同様に、スコア関数‘DF-CP’と‘DP-CP’の比較を行う。再現率を比べると、英日方向、日英方向共に、辞書スコア‘DP’を用いるスコア関数の方が再現率が若干高いことがわかる。これは、表4、表5で示したように、対訳辞書スコア‘DF’よりも対訳辞書スコア‘DP’の方が、正解訳語を生成可能な用語数が多いことによると考えられる。一方、精度に関しては、対訳辞書スコア‘DF’を用いるスコア関数の方が、若干高いことがわかる。F値を見た場合、対訳辞書スコア‘DF’を用いるスコア関数と対訳辞書スコア‘DP’を用いるスコア関数にはほとんど差がない。

また、‘CF’、‘CP’、‘CO’の3つのコーパススコアを用いたスコア関数の再現率を比較すると、‘DF-CF’、‘DF-CP’、‘DF-CO’の間で大きな差はない。ただし、‘DF-CF’と‘DF-CO’を比較すると、‘DF-CO’の再現率及び精度が若干高い。これは、生成された不正解の訳語が一般的な語であった場合、コーパススコア‘CF’では高いスコアが与えられてしまうことが原因と考えられる。これに対して、‘DF-CO’においては、訳語候補のスコアの値が対訳辞書スコア‘DF’の値のみによって決定されるので、コーパス中に高頻度に出現する一般語に対して過剰に高いスコアを与えるということはない。一方、コーパススコア‘CP’に注目すると、コーパススコア‘CP’を用いたスコア関数‘CF-CP’及び‘DP-CP’の精度が他のスコア関数より低いことがわかる。コーパススコア‘CP’を用いると訳語候補全体がコーパスに存在しなくても、スコアが付与されることとなり、不適切な訳語候補が数多く出力されていると考えられる。

上記のコーパススコア‘CP’の評価と関連して、(藤井, 石川 2000)で用いられた確率に基づくスコア関数の評価として、部分対応対訳辞書に加えて英辞郎自体も用いることができるよう拡張を加えた‘DP-CP’に注目する。このスコアは、他のスコア関数と比べ、性能に大きな差はないが、再現率が若干高く、精度が若干低い値となっている。したがって、F値でみると、他のスコア関数とほとんど差がないことがわかる。

対訳辞書スコアのみを用いるスコア関数‘DF’は、英日方向では再現率が他のスコア関数より低いが、日英方向では他のスコア関数に比べて遜色のない結果となった。しかしながら、精度及びF値は他のスコア関数に比べ極めて低い。このことから、訳語候補の順位付けには、コーパスに基づくスコアを用いることが必要であることがわかる。

コーパススコアのみを用いるスコア関数‘CF’及び‘CP’に着目すると、英日方向、日英方向共に、再現率、精度、F値が最も低い。このことより、訳語候補の順位付けには、辞書に生起する頻度に基づく何らかのスコアを利用する必要があることがわかる。

前置詞とハイフンの規則の評価

3.3節の脚注8で述べた前置詞“of”とハイフンの挿入・削除に関する規則の効果について述べる。スコア関数として‘DF-CO’を用いて、これらの規則の評価を行ったところ、英日方向において、この2つの規則を加えることで、正解数が18個増加した。逆に日英方向の場合、この

2つの規則を加えることで、正解数が7個増加した。このことから、この2つの規則は正解数の向上に有効であることがわかる。

分野別の再現率と精度に関する考察

ここでは、分野別に訳語推定の性能を評価する。まず、 Y_S 全体に対する再現率の定義と同様に、各分野に属する用語のうち、スコア1位の訳語候補が正解である用語の割合を分野別再現率と定義する。同様に、各分野に属する用語のうち、訳語候補が1つ以上生成される用語に対して、スコア1位の訳語候補が正解である割合を分野別精度と定義する。

ウェブから収集した専門分野コーパスを用いるスコア関数の中では最もF値の値が高かったスコア関数‘DF-CO’を対象とし、分野別再現率を表8に示す。これと、表4の T_{gc} に属する用語の割合を比較すると、 T_{gc} に属する用語の割合が小さい分野ほど、再現率が低くなっていることがわかる。正解訳語が生成可能な用語を増やすための対訳辞書の強化と、コーパスに正解訳語が含まれる割合の改善が課題となる。

表9に、訳語候補が1つ以上生成される用語に対する分野別精度を示す。スコア関数としては‘DF-CO’を用いた。表8において Y_S 全体に対する分野別再現率が他の分野と比べて低かったのは「化学物質及び薬物」の分野であったが、表9においては、「化学物質及び薬物」の分野の精度は英日方向、日英方向とも80%以上という高い精度となっている。訳語候補が1つ以上生成される用語に対する評価では、 T_{gc} に属する割合との相関はないことがわかる。

表8 スコア関数‘DF-CO’の分野別再現率

分野	英語→日本語		日本語→英語	
	top 1	top 10	top 1	top 10
電磁気学	40%	47%	60%	63%
電気工学	46%	49%	44%	46%
光学	32%	32%	42%	48%
プログラム言語	68%	71%	75%	82%
プログラミング	61%	64%	57%	71%
コンピュータ	32%	32%	38%	43%
解剖学	52%	57%	36%	41%
疾患	48%	49%	50%	52%
化学物質及び薬物	35%	37%	33%	35%
物理化学及び統計学	51%	56%	43%	51%
全体	44.7%	47.6%	43.9%	48.6%

表 9 訳語候補が 1 つ以上生成される用語に対する分野別精度と F 値: スコア関数 ‘DF-CO’ の結果

(a) 英語→日本語

分野	出力あり	1 位正解			10 位以内正解		
		個数	精度	F 値	個数	精度	F 値
電磁気学	16	12	75%	52%	14	88%	61%
電気工学	26	19	73%	57%	20	77%	60%
光学	14	10	71%	44%	10	71%	44%
プログラム言語	25	19	76%	72%	20	80%	75%
プログラミング	21	17	81%	69%	18	86%	73%
コンピュータ	53	32	60%	42%	32	60%	42%
解剖学	64	47	73%	61%	52	81%	67%
疾患	54	41	76%	59%	42	78%	60%
化学物質及び薬物	36	29	81%	48%	31	86%	52%
物理化学及び統計学	70	50	71%	59%	55	79%	65%
全体	379	276	72.8%	55.4%	294	77.6%	59.0%

(b) 日本語→英語

分野	出力あり	1 位正解			10 位以内正解		
		個数	精度	F 値	個数	精度	F 値
電磁気学	21	18	86%	71%	19	90%	75%
電気工学	26	18	69%	54%	19	73%	57%
光学	16	13	81%	55%	15	94%	64%
プログラム言語	25	21	84%	79%	23	92%	87%
プログラミング	22	16	73%	64%	20	91%	80%
コンピュータ	62	38	61%	47%	43	69%	53%
解剖学	51	33	65%	46%	37	73%	52%
疾患	51	43	84%	63%	45	88%	66%
化学物質及び薬物	34	28	82%	47%	29	85%	49%
物理化学及び統計学	61	43	70%	54%	50	82%	63%
全体	369	271	73.4%	55.0%	300	81.3%	60.9%

翻訳ソフトによる翻訳性能との比較

Y_S の用語を市販の翻訳ソフトで翻訳し、その翻訳性能と表 6 の再現率を比較する。翻訳ソフトとしては、富士通の「ATLAS 翻訳パーソナル 2003」、東芝の「The 翻訳オフィス V6.0」、IBM の「インターネット翻訳の王様バイリンガル Version 5」の 3 種類を用いた。この実験は、構成要素の訳語選択がどの程度できるのかを調べるのが目的であるので、翻訳ソフトのオプションの専門用語辞書は使用しなかった。このうち最も性能が良かったのは、富士通の翻訳ソフト

「ATLAS 翻訳パーソナル 2003」で翻訳した場合で、翻訳結果が正解であった用語の割合は英日方向 26.7%，日英方向で 38.1% であった。スコア関数 ‘CF’ 及び ‘CP’ を除くすべてのスコア関数の再現率は、翻訳ソフトによる翻訳結果が正解であった用語の割合を上回っている。

4.3 正解訳語が生成できない原因の分析

本節では、対訳辞書として、英辞郎と部分対応対訳辞書 B_P, B_S を用いたときに、正解訳語が生成できない場合に関して、その主原因を調査した結果について述べる。分析対象は、表 4 に示した生成可能な用語の集合 (T_g) に属さない用語であり、英日方向で 174 語、日英方向で 150 語である。分析結果を表 10 に示す。まず、用語とその正解訳語が構成的に対応しておらず、本手法では扱えないものが、英日方向で 33%，日英方向で 31% 存在した。これは英日方向で、集合 Y_S 全体の 9.4%，日英方向で 7.5% に相当する。次に、辞書にエントリがないことが原因であるものが、英日方向で 42%，日英方向で 44% 存在した。「デジタル」と「デジタル」のような表記の揺れが原因であるものが、英日方向で 3%，日英方向で 10% 存在した。これらに対しては、部分対応対訳辞書の強化が課題となる。英語用語中に前置詞を含まず英語と日本語で語順が異なるものが、英日方向で 10%，日英方向で 9% 存在した。これらは、医学分野に多いため、特定の語が構成要素に現れた場合は、語順の入れ替えを行うという対処法が考えられる。

4.4 スコア関数 ‘DF-CO’ とスコア関数 ‘DF-CF_{f-w}’ の併用と誤り分析

4.2 節の評価では、ウェブから収集した専門分野コーパスを用いる方法は精度に優れ、ウェブ全体を用いる方法は再現率に優れることを示した。そこで、本節では、両者を相補的に統合するために、ウェブから収集した専門分野コーパスを用いるスコア関数で訳語推定を行い、その結果、訳語候補が 1 つも生成されない場合は、サーチエンジンを通してウェブ全体を利用するスコア関数を用いるというアプローチの評価と誤り分析を行う。ここでは、各々の方法におけるスコア関数としては、それぞれ、最も F 値が大きかった ‘DF-CO’、および、‘DF-CF_{f-w}’ を用いる。

まず、スコア関数 ‘DF-CO’ で訳語推定を行った結果を、図 4 で示した生成可能性に関する分類を利用して整理した。その結果を表 11 に示す。4.1 節で説明したように、 T_{gc} は生成可能かつ専門分野コーパスに含まれる正解訳語が存在する用語の集合である。 T_g は、正解訳語を生成可能な用語の集合なので、 $T_g - T_{gc}$ は、いずれかの正解訳語は生成可能であるが、生成可能かつ専門分野コーパスに含まれる正解訳語は存在しない用語の集合となる。そして、 $\overline{T_g}$ は正解訳語が生成不可能な用語の集合である。これらの、 T_{gc} 、 $T_g - T_{gc}$ 、及び $\overline{T_g}$ に含まれる用語を、スコアが 1 位の訳語候補が正解か、不正解か、もしくは、訳語候補が出力されないかによって、それぞれ再分類した。

さらに、 T_{gc} の用語のうち、スコア 1 位の訳語候補が「不正解」のものと「なし」のものに關

表 10 生成不可の原因分析：集合 Y_S のうち、正解訳語が生成不可能な用語を対象

(a) 英語→日本語

生成不可の主原因	個数	割合
非構成的	58	33%
辞書にエントリがない	73	42%
表記の揺れ	6	3%
前置詞による順序交換	6	3%
前置詞なし順序交換	18	10%
訳語に「の」が必要	2	1%
「性」を挿入する必要	1	1%
アルファベットのままにすべき	2	1%
正解訳語に「・」を含む	1	1%
複数形で辞書引き失敗	6	3%
ハイフンの挿入が必要	1	1%
合計	174	100%

(b) 日本語→英語

生成不可の主原因	個数	割合
非構成的	46	31%
辞書にエントリがない	66	44%
表記の揺れ	15	10%
前置詞による順序交換	4	3%
前置詞なし順序交換	13	9%
「性」を外す必要	2	1%
アルファベットのままにすべき	1	1%
用語に「・」を含む	1	1%
正解訳語が複数形	1	1%
訳語中に冠詞が必要	1	1%
合計	150	100%

表 11 スコア関数 DF-CO の訳語推定結果の分析：集合 Y_S を対象

	スコア 1 位の訳語候補					
	英語→日本語			日本語→英語		
	正解	不正解	なし	正解	不正解	なし
T_{gc}	276	30	26	271	36	13
$T_g - T_{gc}$	-	40	71	-	40	107
$\overline{T_g}$	-	33	141	-	22	128
Y_S	276	103	238	271	98	248

して、正解訳語が1位とならなかった原因の分析を行った。英日方向では56個、日英方向では49個がその対象である。その結果を表12に示す。まず、正解訳語の辞書スコアがゼロとなることが原因であるものがあった。これは、辞書スコア‘DF’が、部分対応対訳辞書に一度しか現れない訳語対のスコアをゼロとするように設計されているためである。次に、正解訳語が訳語候補の生成過程で枝刈りされてしまっていることがあった。また、「その他」の理由としては、コーパス中に出現する頻度を考慮したスコア関数を用いていないことなどが挙げられる。

次に、スコア関数‘DF-CO’で、訳語候補が生成されなかった用語を対象として、スコア関数‘DF-CF_{f-w}’を利用して訳語推定を行い、その性能を評価した。対象は、英日方向では238個、日英方向では248個の用語である。評価結果を表13に示す。ここではまず、1つ以上の訳語候補が outputされたか否かにより、「あり」と「なし」に分類している。さらに、それぞれの分類に対し、正解訳語が生成可能か否かによって、さらに分類している。そして、正解訳語が生成可能かつ、訳語候補が outputされる用語に対しては、正解訳語のスコアの順位でさらに分類を行っている。ここで、正解訳語がスコア1位となったものは、英日方向で63個、日英方向で83個である。これと、スコア関数‘DF-CO’の正解を合わせると、正解数は英日方向で339個、日英

表12 スコア関数DF-COにおいて、 T_{gc} 中の用語に対して、正解訳語のスコアが1位とならない原因の分析

正解訳語のスコアが1位とならない主原因	英語→日本語		日本語→英語	
	個数	割合	個数	割合
正解訳語の辞書スコアがゼロ	13	23%	10	20%
正解訳語が生成過程で枝刈りされる	25	45%	10	20%
その他	18	32%	29	59%
合計	56	100%	49	100%

表13 スコア関数DF-COとスコア関数DF-CF_{f-w}の差分の分析：スコア関数DF-COで訳語候補が生成されない用語を対象

		英語→日本語		日本語→英語	
		正解訳語生成可能性		正解訳語生成可能性	
		可	不可	可	不可
訳語候補出力	あり	1位 63	39	1位 83	43
		2~10位 4		2~10位 13	
		10位以下 0		10位以下 1	
		出力されない 8		出力されない 16	
	なし	22	102	7	85
合計		238		248	

方向で 354 個となる。集合 Y_S 全体に対する再現率を求めるとき、英日方向で 54.9%，日英方向で 57.4% となり、他のどのスコア関数よりも高い。また、最終的に訳語候補が 1 つ以上出力されるものを対象にした評価を行うと、精度は、英日方向で 68.8%，日英方向で 67.4% となり、スコア関数 ‘DF-CO’ に比べて精度を大きく下げるところなく、正解数を増やすことに成功した。さらに、F 値は、英日方向で 61.1%，日英方向で 62.0% となり、他のどのスコア関数よりも高い。このことから、スコア関数 ‘DF-CO’ とスコア関数 ‘DF-CF_{f-w}’ を組み合わせるアプローチが有効であることがわかる。

ここで、正解訳語が 2 位以下に出力される用語に対しては、(木田, 外池, 宇津呂, 佐藤 2006) で提案されている用語の分野判定の技術により訳語候補の分野判定を行い、分野外の訳語候補を削除することによって、正解訳語を 1 位にできる可能性があると考えられる。また、正解訳語が生成不可で訳語候補が 1 つ以上出力されている用語に対しては、訳語候補の分野判定を行うことによって、候補数をゼロにできる可能性がある。

5 関連研究

関連研究として、(藤井, 石川 2000) は、言語横断情報検索の目的のために要素合成法による訳語推定法を提案した。本論文では、ここで提案されているスコア関数に対して、部分対応対訳辞書だけでなく、英辞郎自体も利用できるように拡張し（スコア関数 ‘DP-CP’），他のスコア関数との比較を行った。その結果、スコア関数 ‘DP-CP’ は他のスコア関数と比べ、 Y_S 全体に対する評価では最も再現率が高かったが、不正解訳語も多く生成されるため、訳語候補が 1 つ以上生成される用語に対する評価では、精度は高くなかった。そして、F 値に関しては、他のスコア関数とほとんど差がなかった。(藤井, 石川 2000) の手法と、本論文で提案した手法の重要な違いの一つは、(藤井, 石川 2000)においては、訳語推定対象の用語が属する分野の文書のみを含むコーパスではなく、様々な専門分野にわたる 65 種類の日本の学会から出版された技術論文を集めたものをコーパスとして利用していることである。また、(藤井, 石川 2000)において、彼らは言語横断情報検索の性能のみを評価し、訳語推定の性能評価はしていない。

(阿玉, 橋本, 徳永, 田中 2004) も、言語横断情報検索の性能を評価対象として、クエリー翻訳の方法を提案している。この研究では、コーパススコアは NTCIR-1(Kando, Kuriyama, and Nozue 1999)，または、NTCIR-2(Kando, Kuriyama, and Yoshioka 2001) の言語横断情報検索タスクの検索課題文書（論文等の技術文書）から求める。コーパススコアには、(藤井, 石川 2000) で提案されたスコアと合わせて χ^2 検定を用いたスコアを併用している。しかしながら、2 つのコーパススコアの併用は、言語横断情報検索の精度向上には貢献しなかったと報告されている。また、カタカナ語に関しては翻字技術を適用している。

(Baldwin and Tanaka 2004) も要素合成法による訳語推定手法を提案している。コーパスに基

づく 8 つの素性と辞書に基づく 6 つの素性とテンプレートに基づく 2 つの素性を立て、SVM を利用して訳語候補のスコア関数を学習している。この論文でも、辞書に基づく素性でのみ、もしくは、コーパスに基づく素性でのみスコア関数を構成するよりも、両者を利用した方が精度が良いことが報告されている。(Baldwin and Tanaka 2004) の手法と、本論文で提案した手法の重要な違いは、(Baldwin and Tanaka 2004) においては、コーパスとして、英語側では Reuters Corpus を、日本語側では毎日新聞を利用しているのに対し、本論文では、専門分野コーパスを利用した場合と、サーチエンジンを通してウェブ全体を利用した場合の比較を行っている。また、(Baldwin and Tanaka 2004) においては、訳語推定対象の用語を、英語 2 単語または、日本語 2 形態素のものに限定している。

(Cao and Li 2002) もまた、複合語に対する要素合成法による訳語推定法を提案した。(Cao and Li 2002) の手法では、用語の訳語候補は、用語の構成要素の訳語を結合することによって構成的に生成され、サーチエンジンを通してウェブ全体を用いて検証される。本論文では、サーチエンジンを通してウェブ全体を用いて訳語候補の検証をするスコア関数を導入することによって、(Cao and Li 2002) で提案されたアプローチの評価を行い、ウェブから収集した専門分野コーパスを用いる方法と比較した。その結果、訳語候補が一つ以上出力される場合においては、サーチエンジンを通してウェブ全体を用いるよりも、ウェブから収集した専門分野コーパスを用いる方が精度が良いことがわかった。その一方で、再現率に優れるのは、サーチエンジンを通してウェブ全体を用いる方法であった。そこで、この 2 つの方法の長所を生かすために、まず、ウェブから収集した専門分野コーパスを用いる方法で訳語推定を行い、訳語候補が一つも得られなかった場合、サーチエンジンを通してウェブ全体を用いて訳語推定を行う方法を評価した。その結果、本論文で評価したどのスコア関数よりも高い F 値を達成できることができた。なお、(Cao and Li 2002) においては、英語の用語に対して中国語の訳語を推定しているが、訳語推定対象の用語は英語 2 単語から構成されるものに限定されている。

(前田、吉川、植村 2000) は、言語横断情報検索のためのクエリー翻訳の方法を提案している。まず、要素合成法によりクエリーの訳語候補を生成する。次に、ウェブ上の頻度が一定数以上の訳語候補に対して、それぞれの訳語候補のスコアは、訳語候補の構成要素間の相互情報量を拡張した尺度で計算される。最後に、スコアの閾値を越える訳語候補を（サーチエンジンの）OR 演算子で結合したものを、クエリーの翻訳結果とする。評価は言語横断情報検索の性能に関して行っているので、訳語推定結果の比較はできないが、(前田他 2000) らの手法は、本論文で言えば、コーパススコアのみのスコア関数を利用した手法に対応する。

(木村、前田、宮崎、吉川、植村 2004) も、言語横断情報検索のために、クエリー翻訳における訳語曖昧性解消の方法を提案している。準備として、あらかじめ Yahoo! の日英のウェブディレクトリのそれぞれのカテゴリにおいて、特徴語の抽出と重み付与をし、日英のカテゴリの対応付けをしてしておく。検索をするときは、まず、クエリーに含まれる単語とカテゴリの特徴

語を利用して適合するカテゴリを決める。次に、クエリーに含まれる各単語に対して、訳語を対訳辞書で調べる。そして、適合カテゴリの特徴語となっている訳語のうち、特徴語の重みが最も大きいものをその単語の訳語と決定する。

6 おわりに

本論文では、ウェブを利用した専門用語の訳語推定法について述べた。これまでに行われてきた訳語推定の方法の1つに、パラレルコーパス・コンパラブルコーパスを用いた訳語推定法があるが、既存のコーパスが利用できる分野は極めて限られている。そこで、本論文では、訳を知りたい用語を構成する単語・形態素の訳語を既存の対訳辞書から求め、これらを結合することにより訳語候補を生成し、単言語コーパスを用いて訳語候補を検証するという手法を採用した。しかしながら、単言語コーパスであっても、研究利用可能なコーパスが整備されている分野は限られている。このため、本論文では、ウェブをコーパスとして用いた。ウェブを訳語候補の検証に利用する場合、サーチエンジンを通してウェブ全体を利用する方法と、訳語推定の前にあらかじめ、ウェブから専門分野コーパスを収集しておく方法が考えられる。本論文では、評価実験を通して、この2つのアプローチを比較し、その得失を論じた。また、訳語候補のスコア関数として多様な関数を定式化し、訳語推定の性能との間の相関を評価した。実験の結果、ウェブから収集した専門分野コーパスを用いた場合、ウェブ全体を用いるよりカバレージは低くなるが、その分野の文書のみを利用して訳語候補の検証を行うため、誤った訳語候補の生成を抑える効果が確認され、高い精度を達成できることがわかった。また、ウェブ全体を用いる方法とウェブから収集した専門分野コーパスを用いる方法を相補的に結合することにより、再現率とF値を改善できることを示した。

今後の課題として、訳語推定対象の分野の用語を十分に含むできるだけ小さいコーパスを収集することが挙げられる。また、本論文で提案した、ウェブを用いた要素合成法による訳語推定法を、他の訳語推定技術と相補的に用いることが挙げられる。相補的な技術としては、用語とその訳語が併記されたテキストの利用 (Nagata, Saito, and Suzuki 2001; Huang, Zhang, and Vogel 2005) や、固有名詞の翻字の技術 (Knight and Graehl 1998; Oh and Choi 2005) などが挙げられる。また、用語の分野判定の技術 (木田他 2006) を利用することにより、不適切な訳語候補を削除することが挙げられる。応用的な課題としては、本論文で提案した専門用語の訳語推定手法を、例えば、ウェブからの関連語収集手法 (佐々木, 宇津呂, 佐藤 2006) や、論文からの用語抽出 (馬場, 外池, 宇津呂, 佐藤 2006) の結果に対して適用することが考えられる。

参考文献

- 阿玉泰宗, 橋本泰一, 徳永健伸, 田中穂積 (2004). “日英言語横断情報検索のための翻訳知識の獲得.” 情報処理学会論文誌：データベース, **45** (SIG10(TOD23)), pp. 37–48.
- Baldwin, T. and Tanaka, T. (2004). “Translation by Machine of Compound Nominals: Getting it Right.” In *Proc. ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pp. 24–31.
- 馬場康夫, 外池昌嗣, 宇津呂武仁, 佐藤理史 (2006). “対訳辞書とウェブを利用した専門文書中の用語の訳語推定.” 言語処理学会第12回年次大会論文集, pp. 416–419.
- Cao, Y. and Li, H. (2002). “Base Noun Phrase Translation Using Web Data and the EM Algorithm.” In *Proc. 19th COLING*, pp. 127–133.
- コンピュータ用語辞典編集委員会(編) (2001). 英和コンピュータ用語大辞典. 日外アソシエーツ.
- 藤井敦, 石川徹也 (2000). “技術文書を対象とした言語横断情報検索のための複合語翻訳.” 情報処理学会論文誌, **41** (4), pp. 1038–1045.
- Fung, P. and Yee, L. Y. (1998). “An IR Approach for Translating New Words from Nonparallel, Comparable Texts.” In *Proc. 17th COLING and 36th ACL*, pp. 414–420.
- Huang, F., Zhang, Y., and Vogel, S. (2005). “Mining Key Phrase Translations from Web Corpora.” In *Proc. HLT/EMNLP*, pp. 483–490.
- 医学用電子化AI辞書研究会(編) (1996). 25万語医学用語大辞典. 日外アソシエーツ.
- Kando, N., Kuriyama, K., and Yoshioka, M. (2001). “Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop.” In *Proc. 2nd NTCIR Workshop Meeting*, pp. 73–96.
- Kando, N., Kuriyama, K., and Nozue, T. (1999). “NACESIS test collection workshop (NTCIR-1).” In *Proc. 22nd SIGIR*, pp. 299–300.
- 木田充洋, 外池昌嗣, 宇津呂武仁, 佐藤理史 (2006). “ウェブを利用した専門用語の分野判定.” 電子情報通信学会論文誌, **J89-D** (未定).
- 木村文則, 前田亮, 宮崎純, 吉川正俊, 植村俊亮 (2004). “Webディレクトリを言語資源として利用した言語横断情報検索.” 情報処理学会論文誌：データベース, **45** (SIG7(TOD22)), pp. 208–217.
- Knight, K. and Graehl, J. (1998). “Machine Transliteration.” *Computational Linguistics*, **24** (4), pp. 599–612.
- 前田亮, 吉川正俊, 植村俊亮 (2000). “言語横断情報検索におけるWeb文書群による訳語曖昧性解消.” 情報処理学会論文誌：データベース, **41** (SIG6(TOD7)), pp. 12–21.
- Matsumoto, Y. and Utsuro, T. (2000). “Lexical Knowledge Acquisition.” In Dale, R., Moisl, H.,

- and Somers, H. (Eds.), *Handbook of Natural Language Processing*, chap. 24, pp. 563–610. Marcel Dekker Inc.
- マグローヒル科学技術用語大辞典編集委員会（編）（1998）。マグローヒル科学技術用語大辞典。日刊工業新聞社。
- 長尾真, 石田晴久, 稲垣康善, 田中英彦, 辻井潤一, 所真理雄, 中田育男, 米澤明憲（編）（1990）。岩波情報科学辞典。岩波書店。
- Nagata, M., Saito, T., and Suzuki, K. (2001). “Using the Web as a Bilingual Dictionary.” In *Proc. Workshop on Data-driven Methods in Machine Translation*, pp. 95–102.
- Oh, J. and Choi, K. (2005). “Automatic Extraction of English-Korean Translations for Constituents of Technical Terms.” In *Proc. 2nd IJCNLP*, pp. 450–461.
- Rapp, R. (1999). “Automatic Identification of Word Translations from Unrelated English and German Corpora.” In *Proc. 37th ACL*, pp. 519–526.
- 佐々木靖弘, 宇津呂武仁, 佐藤理史 (2006). “関連用語収集問題とその解法。”自然言語処理, 13(3), pp. 151–175.
- 高木俊宏, 木田充洋, 外池昌嗣, 佐々木靖弘, 日野浩平, 宇津呂武仁, 佐藤理史 (2005). “ウェブを利用した専門用語対訳集自動生成のための訳語候補収集。”言語処理学会第11回年次大会論文集, pp. 13–16.
- Tonoike, M., Kida, M., Takagi, T., Sasaki, Y., Utsuro, T., and Sato, S. (2005). “Effect of Domain-Specific Corpus in Compositional Translation Estimation for Technical Terms.” In *Proc. 2nd IJCNLP, Companion Volume*, pp. 116–121.

略歴

外池 昌嗣：2001年京都大学工学部情報学科卒業。2003年同大学大学院情報学研究科修士課程知能情報学専攻修了。2007年同大学大学院情報学研究科博士後期課程修了予定。自然言語処理の研究に従事。

宇津呂武仁：1989年京都大学工学部電気工学第二学科卒業。1994年同大学大学院工学研究科博士課程電気工学第二専攻修了。京都大学博士（工学）。奈良先端科学技術大学院大学情報科学研究科助手、豊橋技術科学大学工学部情報工学系講師、京都大学情報学研究科知能情報学専攻講師を経て、2006年より筑波大学大学院システム情報工学研究科知能機能システム専攻助教授。自然言語処理の研究に従事。

佐藤 理史：1983年京都大学工学部電気工学第二学科卒業。1988年同大学院博士課程研究指導認定退学。京都大学工学部助手、北陸先端科学技術大学院大学情報科学研究科助教授、京都大学情報学研究科助教授を経て、2005年より

名古屋大学大学院工学研究科教授。工学博士。自然言語処理、情報の自動編集等の研究に従事。

(2006年8月20日 受付)

(2006年10月24日 再受付)

(2006年11月24日 採録)